Demo报告讨论 20250415

数据上传部分

1. 上传文件类型

- [ ] 上传的是搜库后的peaktable，如果按目前的搜库方案，在搜库阶段就会有group表，所以这里其实可以省去down下来重新传（客户自己提供peaktable除外），目前直接传搜库的group表还是很丝滑。
- [ ] 比较的组别，是否可以选择多个比较方案？比如项目有ABC三个组，A是control，BC是相同突变体不同处理，需要对比AB，AC，BC，是否可以在这里同时选择？因为要重新对比其他组合，还要重新再来一遍传数据流程。
- [ ] 报错位置：在传完POS的peaktable和group后，再传NEG，会容易出现POS表混乱的情况，只能刷新网页重新上传。

2. 数据预处理

这块内容，建议根据我们目前的peaktable和计算过程来填充默认值，客制化选择保留

- [ ] 缺失值类型：0
- [ ] 填充方式：均值
- [ ] 归一化方法：SUM
- [ ] RSD取log：None

结果部分

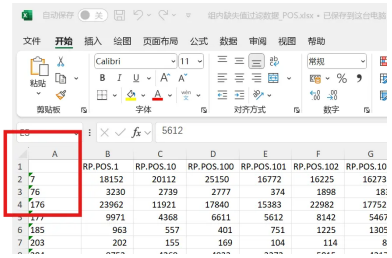
1. 数据总览（网页）

- [ ] 这里缺少一个信息，即peaktable的feature数和鉴定统计，只放了直接对比的，比如想知道这个项目正离子和负离子一共打到多少个feature，多少个能搜到名字的。
- [ ] 这里的raw_feature_num和remove_missing_num没有变化，和后面的preprocess_feature_num有啥区别？

2. 数据预处理

- [ ] （网页）目前点击这里的链接是提示404

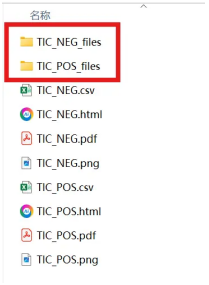
- [ ] 在下载的results表格里，每个表的第一列缺少表头，这里用的是什么ID?



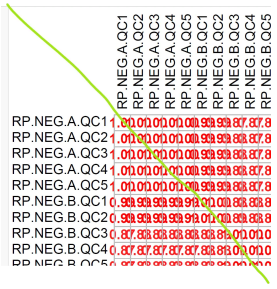
- [ ] 数据预处理提供4个文件夹，这3个里面表格缺失

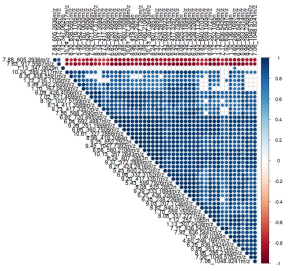
3. QC质控

- [ ] 首先针对所有results（down下来的），建议按分析顺序编号，比如1.TIC，2.corr，3. RSD，4. PCA，5. PLD-DA。其他文件夹也是如此。
- [ ] results文件，TIC里面这两个文件夹是tmp文件？可以不要放在结果里。

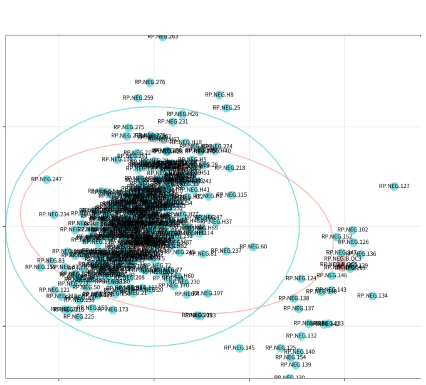


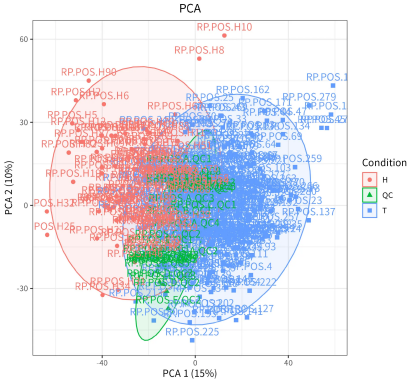
- [ ] corr，方框里建议不要显示数字，换成圆圈，另外整张图可以换成对角线切割，不用显示正方形，这样看起来不会很复杂





- [ ] PCA有个问题，如果把所有样本归成一个组，再来算跟QC的PCA，跟按实际组来分后一起算的结果会有不同，建议直接分组计算，不然QC差异会被拉得很大，比如后面多元统计的这个PAC图，同样的数据，差异大小完全不同。反而多元统计的时候，不用放QC，只需要放样本分组的数据。





- [ ] PCA的图，如果这里只是要凸显QC的离散程度，样本不用打名字，不然根本看不清：
- [ ] QC的PLS-DA，我这边不太明白为啥要做，因为给不了什么参考信息

☐ QC和整体数据质量，建议增加一张图，可作为数据质量控制依据：

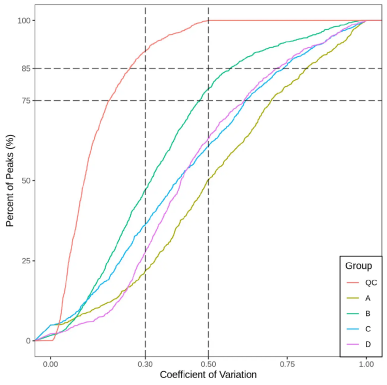
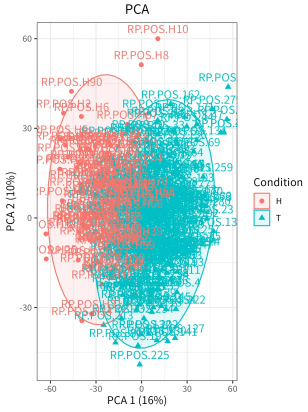


图 10: 各组样本 CV 分布图

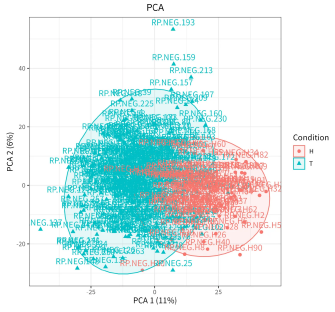
注：横坐标代表 CV 值，纵坐标表示小于对应 CV 值的物质数目占总物质数的比例，不同颜色代表不同的分组样本，QC 为质控样本，其中与 X 轴垂直的两条参考线对应的 CV 值为 0.3 和 0.5，与 X 轴平行的两条参考线对应物质数目占总物质数的 75% 和 85%。

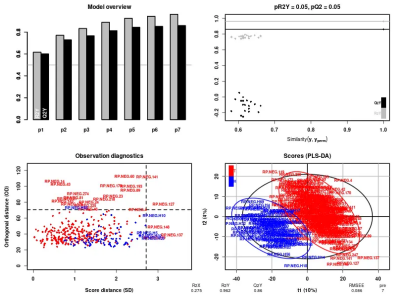
4. 多元统计分析

☐ 组间对比PCA，首先建议在图里标注好POS or NEG，因为所有legend都一样，区分不了，其次这个图的比例可否调整为正方形？视觉上纵坐标会差异很大，其实纵坐标权重反而小；另外，点的名字，如果在置信区间里面是否可以不写？只标注离群点的名字即可：



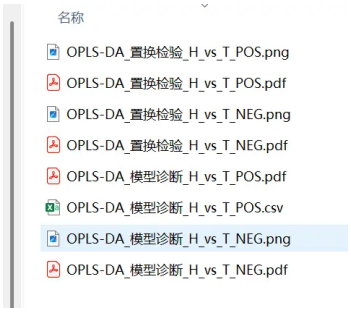
☐ PLS-DA，1. 首先同样建议图里面标好POS or NEG；2. 去除QC组，QC只放在QC部分，其他后面所有分析不要带；3. 结果里面的PLS-DA_H_vs_T_POS，依然是PCA结果，并不是PLS-DA结果，并且名字直接标的PCA；4. 需要的是，PLS-DA建模后，右下角的图，数据提取出来，重新画。





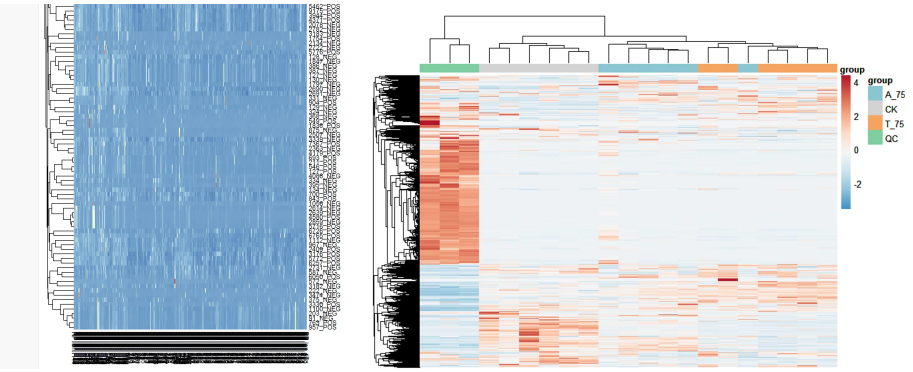
☐ PLS-DA诊断模型，少文件，没有POS的诊断模型png；

☐ OPLS-DA文件夹，少文件，不全；跟前面一样，不要放QC：



5. 差异筛选

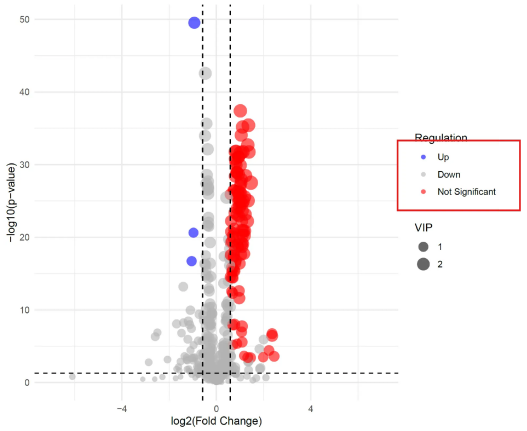
- [ ] HCA：这里作图的数据是什么数据？有做z-score运算吗？
- [ ] HCA：不建议显示很纵坐标名，另外做的时候需要分组：



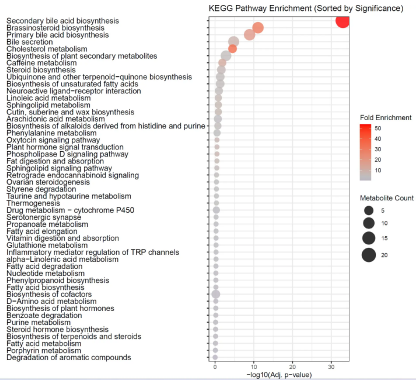
- [ ] HCA: HCA_heatmap_metabolite是什么？如果是用有name的，首先统一命名方式，其次，热图不用写名字，会给表的，而且name这里的合并逻辑是？peaktable应该是一个feature对应一个name的可能存在不同feature对应1个name，但没有一个feature对应多个name：

	A	B	C	D
1		RP.ion.H1	RP.ion.H10	RP.ion.H11
2	2096_NEG_Glycocholic acid;Glyco-gamma-muricholic acid;Glyco-beta-muricholic acid	-0.1981716	-0.1977052	-0.2566108
3	2709_NEG_Tauro-beta-muricholic acid;Tauro-gamma-muricholic acid;Taurocholic acid	-0.1855516	0.02182371	-0.1821633
4	1936_NEG_Glycodeoxycholic acid;Glycochenodeoxycholic acid	-0.3818028	-0.5865649	-0.1570776
5	4367_POS_Glycochenodeoxycholic acid;Glycodeoxycholic acid	-0.464802	-0.6749387	-0.1478223
6	2540_NEG_Taurochenodeoxycholic acid;Taurodeoxycholic acid;Taurohyodeoxycholic acid (THDCA)	-0.3961696	-0.1606353	-0.2865981
7	5241_POS_Taurohyodeoxycholic acid (THDCA);Taurodeoxycholic acid;Taurochenodeoxycholic acid	-0.3418401	-0.2578456	-0.3528045
8	2098_NEG_Glycocholic acid;Glyco-gamma-muricholic acid;Glyco-beta-muricholic acid	-0.6259822	0.06819966	-0.0342954

- [ ] corr：1. feature上次说了不用全放，只放系数最大的前50或者前20即可，其他的数据有提供表格；2. 带name的（文件名要统一）也是，大于20的放20，小于20的放全部；
- [ ] corr： sample的相关性是啥意思？这个是怎么算的？
- [ ] VIP：文件夹里面的表格，只给了VIP>1且有name的结果，这里不对，应该是所有feature的VIP值表；同样，distribution数量统计，需要统计所有feature的VIP分布
- [ ] volcano: 1. 少VIP筛选标准的火山图，这个是跟提交的筛选标准有关还是就不用？2. 火山图的legend错误。



- [ ] 差异代谢物表达量，这里面的两个图都没看明白
- [ ] 差异代谢物网络图：需要提供所有feature的，和有name的化合物的，目前只画了有name的
- [ ] KEGG富集：1. 气泡图，建议调整画图用参数，把很坐标改成fold enrichment，颜色用-log10P，大小为化合物数量；2. 通路名右对齐。


KEGG Pathway Enrichment (Sorted by Significance)
Secondary bile acid biosynthesis
Brassinosteroid biosynthesis
Primary bile acid biosynthesis
Bile secretion
Cholesterol metabolism
Biosynthesis of plant secondary metabolites
Caffeine metabolism
Steroid biosynthesis
Ubiquinone and other terpenoid−quinone biosynthesis
Biosynthesis of unsaturated fatty acids
Neuroactive ligand−receptor interaction
Linoleic acid metabolism
Sphingolipid metabolism
Cutin, suberine and wax biosynthesis
Arachidonic acid metabolism
Biosynthesis of alkaloids derived from histidine and purine
Phenylalanine metabolism
Oxytocin signaling pathway
Plant hormone signal transduction
Phospholipase D signaling pathway
Fat digestion and absorption
Sphingolipid signaling pathway
Retrograde endocannabinoid signaling
Ovarian steroidogenesis
Styrene degradation
Taurine and hypotaurine metabolism
Thermogenesis
Drug metabolism − cytochrome P450
Serotonergic synapse
Propanoate metabolism
Fatty acid elongation
Vitamin digestion and absorption
Glutathione metabolism
Inflammatory mediator regulation of TRP channels
alpha−Linolenic acid metabolism
Fatty acid degradation
Nucleotide metabolism
Phenylpropanoid biosynthesis
Fatty acid biosynthesis
Biosynthesis of cofactors
D−Amino acid metabolism
Biosynthesis of plant hormones
Benzoate degradation
Purine metabolism
Steroid hormone biosynthesis
Biosynthesis of terpenoids and steroids
Fatty acid metabolism
Porphyrin metabolism
Degradation of aromatic compounds
0
10
20
30
−log10(Adj. p−value)
Fold Enrichment
50
40
30
20
10
Metabolite Count
5
10
15
20

☐ 富集，增加MESA分析